DOCUMENT RESUME

ED 270 463 TM 860 341

AUTHOR Hickman, Linda J.
TITLE Freedom from the Tyranny of the Campus Main-Frame: Handling the Statistical Analysis of a 10-year Survey Research Study with a Personal Computer.
PUB DATE Apr 86
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cost Effectiveness; Data Analysis; Data Processing; Evaluation Criteria; Higher Education; Microcomputers; *Programing; Research Methodology; *Research Needs; *Research Problems; Statistical Analysis; *Surveys
IDENTIFIERS *CRISP Computer Program; *Statistical Packages

ABSTRACT

Technological advances in microcomputer hardware and software, including size of memory and increasingly more sophisticated statistical application packages, create a new era in educational research. The alternative to costly main-frame computer data processing and statistical analysis is explored in this paper. In the first section, typical problems of using main-frame computers for survey research are outlined. The second section presents search criteria, including technical tests for accuracy, for selecting a microcomputer statistical package to meet the particular needs of the researcher. Reviews and references for over 30 statistical packages are summarized. The third section presents the selection of and use of CRISP as the solution to the search for a package to handle the data from a 10-year survey research study. In the final section, the impact of changing technology on survey research design and methodology is summarized. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270463

FREEDOM FROM THE TYRANNY OF THE CAMPUS MAIN-FRAME:

Handling the Statistical Analysis of a 10-year Survey Research Study with a Personal Computer

Presented
at the

1986 AERA ANNUAL MEETING
San Francisco
April 1986

by

Linda J. Hickman, Ph.D.
Faculty, Business Analysis
and Computing Systems
School of Business
San Francisco State University
San Francisco, CA 94132

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. J. Hickman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 341

ABSTRACT

FREEDOM FROM THE TYRANNY OF THE CAMPUS MAIN-FRAME:
Handling the Statistical Analysis of a 10-year Survey Research Study with a Personal Computer

Technological advances in microcomputer hardware and software, i.e., size of memory and increasingly more sophisticated statistical application packages, create a new era in educational research. The alternative to costly main-frame computer data processing and statistical analysis is explored in the paper. In the first section, typical problems of using main-frame computers for survey research are outlined. The second section presents search criteria, including technical tests for accuracy, for selecting a microcomputer statistical package to meet the particular needs of the researcher. Reviews and references for over 30 statistical packages are summarized. The third section presents the selection of and use of CRISP as the solution to our own search for a package to handle our ten-year survey research study. In the final section, the impact of changing technology on survey research design and methodology is summarized.

Presented by: Linda J. Hickman, Ph.D., faculty, Business Analysis and Computing Systems, School of Business, San Francisco State University, San Francisco, CA 94132.

Presented at: The 1986 AERA Annual Meeting, San Francisco, CA. April 1986.

KEY WORDS: survey research, statistical application packages

FREEDOM FROM THE TYRANNY OF THE CAMPUS MAIN-FRAME:
Handling the Statistical Analysis of a 10-year Survey Research Study with a Personal Computer

Every educator faces the pressure to complete scholarly research within time and budget constraints. For many years, a large research grant often has been required for survey research especially for costly computerized data processing. More critically, research design and methodology are frequently affected by budget constraints which limit the size and scope of data collection and the amount of analysis that are affordable.

Technological advances in microcomputer hardware and software, i.e., size of memory and increasingly more sophisticated statistical application packages, create a new era in educational research. The alternative to costly main-frame computer data processing and statistical analysis is explored in the following paper. In the first section, typical problems of using main-frame computers for survey research are outlined. The second section presents search criteria, including technical tests for accuracy, for selecting a microcomputer statistical package to meet the particular needs of the researcher. Reviews of over 30 statistical packages are summarized. The third section presents the selection and use of CRISP as the solution to our own search for a package to handle our ten-year survey research study. In the final section, the impact of changing technology on survey research design and methodology is summarized.

I.SURVEY RESEARCH AND MAIN-FRAME COMPUTING CONSTRAINTS

Practical considerations of time and costs affect design and methodological decisions in survey research. Foremost, the size of the survey sample is frequently dictated by the costs of data collection, processing and analysis. As a research statistician, I am often uncomfortable with small samples, which might meet the statistical assumptions of the Central Limit Theorem, but which also create wide confidence intervals! Large data sets, however, can be expensive to process and analyze using a main-frame computer. As the sample size increases, costs for data entry, data storage and runs for data analysis also increase. Ideally, the researcher should be able to select the sample size most appropriate for the confidence level desired. If cost factors related to sample size are reduced, the researcher gains more control over research design. With microcomputing, the costs for data storage and data analysis are not related to the number of observations. A survey file of 10 or over 1,000 will be

stored on a disk at the same cost, and the cost for analysis will only be related to the original cost of the analysis package not to the number of observations processed. In comparison, in a typical academic main-frame computing system, charges are based on quantity: CPU's used, number of tapes stored, number of cards read or scanned, number of pages of output printed, number of hours connected through a modem, etc. Those are general costs, there are always many other costs, such as tape mounting, or off-line plotting and other specialized services. To reduce costs in the typical main-frame system, one must reduce quantity of observations to be processed, number of processing runs, and number of pages of output.

A second problem encountered in utilizing main-frame computing for survey research is losing control over access to the data. When the data is stored in a file in the main-frame,on a disk or tape, or less frequently now, on cards, access is dependent upon programming staff, terminal accessibility, and system status. All of which, also, have costs. Statistical packages for the main-frame which are widely used, such as SPSS and SAS, are command-driven programs which require programming expertise for both creating data files as well as data analysis. Project budgets for survey research using main-frame statistical packages typically include expenses for programming staff. Using a microcomputer package provides two advantages: unlimited access to data with portable data disks and elimination of programming support staff with "user friendly" menu-driven programs.

A related problem of losing control over access to the data is losing control over time scheduling. A survey research project time-line when using the main-frame must take into consideration the vagaries of the system. All main-frame systems routinely schedule "downtime" for system maintenance activities, on average at least four hours per week. While these are sometimes scheduled for non-peak hours (early hours on weekends), nevertheless the system is unavailable at those times. Unscheduled "downtime" is always a possibility which can create havoc in a project schedule. Microcomputers, in contrast, are always available.

Costs for computer charges are closely related to time schedules. In most systems costs are directly linked to the time of use. For example, the costs at one major research university for CPU time range from $.62/CPU/second during peak times (daytime on weekdays) to $1.00/hour for batch processing at off-peak times. Charges for the printer are also linked to time of use. Again, to reduce costs using a main-frame system, the researcher must schedule processing and output in conformance to a system imposed time schedule rather than the researcher's own schedule.

Aware of the constraints imposed by the main-frame computer system, alternatives were sought for processing our ten-year survey data. A brief description of the ten-year study will provide a basis for detailing our statistical analysis requirements and the search for satisfactory solutions. The survey research was designed as a comparative study of business

students' and businessmen's attitudes towards ethical issues in business. The survey instrument replicates the one used in Raymond Baumhart's HARVARD BUSINESS REVIEW study (1961), "How Ethical are Businessmen." Baumhart randomly sampled HBR readers and analyzed 1,512 responses to a questionnaire measuring ethical attitudes. Fifteen years later, Brenner and Molander (1976) again surveyed HBR readers using selected parallel questions and obtained 1,227 responses. Our ten-year study (1975-1985) provides a comparison of student attitudes with those of business practitioners.(1)

The methodology of the ten-year study was straightforward. Student populations, rather than random samples, were used. The survey instrument which contains 53 variables and 5 demograpahic variables was administered to 36 sections of undergraduate and 13 sections of graduate students enrolled in selected Business courses. Over the 10 year time period, students at six institutions in five geographic regions: the northeast, far west, south, mid-atlantic and southwest, were surveyed. The study design also included pre-and post-surveys for some courses in order to provide a possible measure of change in attitudes towards ethical issues upon completion of coursework in business ethics.

An important methodological issue arose which was related to the analysis of observations over time. Since the student populations were drawn from diverse geographic regions over a ten-year time period, there was concern that statistical differences between groups might be falsely ascribed as geographically related or time related, thus, the second phase of the study, the national study, was designed to collect data from all geographic regions in one time period. The national study created another large data set of n>1,500.

While methodological issues regarding the design of a ten-year survey study are of interest, the focus for this presentation is on the management of extensive empirical research using the current technology available to researchers with personal computers. It is now feasible for scholarly research to be conducted using a respectably large data base without the need for major underwriting for the data processing! After many years of depending on campus main-frame computers and their auxiliary programming staff, powerful statistical analysis packages have become available for the micrcomputers. When we began our search several years ago, the scene was quite different. At times the search was similar to opening Pandora's Box -- the unusual constraints and myriad differences of packages appeared chaotic. Using the search criteria outlined below, a comparison of available packages yielded several alternatives which met our requirements.

II. TECHNICAL AND GENERAL CRITERIA FOR SELECTING STATISTICAL PACKAGES FOR THE MICROCOMPUTER.

Individual research needs govern the selection process for determining the choice of a statistical package. There is considerable variation between statistical packages which makes direct comparisons difficult.

Reviews of statistical packages often emphasize different features. Lachenbruch (1983), and Platt & Platt (1981) emphasized accuracy of results, while Lee (1985) focussed more on ease of use. A comparison of statistical programs and graphic features offered in a package as well as general software criteria (ease of use, documentation and support) was used by Fridlund (1985). Similar criteria, i.e., data manipulation, statistical procedures available, output from statistical procedures, ease of use, and documentation was the basis for evaluation of five packages reviewed by Goodban and Hakuta (1984). A comprehensive set of criteria was provided in a review of 24 packages by Carpenter, et al, (1984). In a recent article, Bannister (1985), quotes Paul Tukey, the well-known statistician, as stressing the importance of checking specific criteria such as speed of the package, system overhead, capability to read and write ASCII files, support of graphics, plotters and printers, and accuracy of algorithms and results.

There are both technical and general criteria for evaluating statistical software which provide guidelines in a search for a satisfactory package. Technical criteria are related to statistical factors such as accuracy of algorithms and results, data manipulation utilities, statistical procedures and tests available (parametric and nonparametric) and file building/handling utitlities. General criteria include the user computing environment requirements (hardware and budget considerations) and generalized software criteria such as ease of handling, documentation, and user support.

In our case, there were specific computing environment and data handling needs as well as generalized software concerns which governed our search for a statistical package:

HARDWARE REQUIREMENTS: Initially our system was an IBM-PC with dual disk drives which required a memory expansion board to test some statistical packages. Eventually we installed a hard disk to increase speed and facilitate file handling.

COST: Seeking a medium-priced package (between $200-500).

APPLICATONS/SCOPE: The statistical procedures required were: Descriptive statistics, frequencies, correlation, regression (including step-wise regression), crosstabs, chi-square, ANOVA, scatterplots, histograms, t-test and nonparametric tests.

The file size and variable size needed to be sufficient to handle up to 1,500 observations with at least 60 variables.

EASE OF USE: In addition to our own use of the program, we were seeking a statistical program which could be used by my students in a college-level statistics course. Therefore, a menu-driven program and other ease of use features were part of our goal. Documentation and error handling (error messages within the program's operation) and continuing telephone support for the software were considerations.

ACCURACY: Evidence of testing for statistical accuracy of the program. Were standards, such as those used for testing main-frame statistical computing packages with the Longley data, or specific statements of accuracy of results described in the product literature.

A.TECHNICAL CRITERIA

1. ACCURACY

In a statistical package, numerical accuracy is essential. Main-frame standards for computing accuracy are now being applied to microcomputer software. Unfortunately, not all packages incorporate rigorous standards. While it is beyond the scope of this paper to discuss all of the issues related to accuracy of algorithms and results, an overview of some of the problems and tests for evaluating accuracy are presented. Typical problems involve the regression and standard deviation algorithms used as well as the handling of significant places for result values.

As several reviews have shown, Lachenbruch (1983), Platt and Platt (1981), Fridlund (1985), multiple regression algorithms used in microcomputing statistical packages are a potential source of error depending upon the data used. When independent variables are highly correlated and X variables are closely related, a quantity close to 0 is used as a divisor. The computation must then be very accurate to avoid a serious error in the solution.In an article in 1967, Longley presented a data set with such highly correlated variables. The data described the U.S.population from 1947 and 1962 and was used to predict employment. It was found

that any one of four variables can be used to accurately predict the employment variable and are highly interrelated. Using the Longley data, reviewers have obtained varying results with different packages. (If you wish to use the data for your own testing, refer to a standard statistics text, for example, Chambers, which cites the Longley data.) (2)

Lachenbruch tested six microcomputer packages and two mainframe packages which yielded varying results for accuracy. For example, Longley's correlation of GNP to employment gave a result of -.035819. In comparison, the results in the eight programs ranged from -.035819 to -.03477. The least accurate results were obtained using AIDA, which was only accurate to 1,2 or 3 digits, depending upon the correlation variable. One package, HSD-Regress, did not calculate the regression coefficients giving a warning that the data did not permit it to compute the coefficients accurately. (3)

Maindonald (1984) also provides some checks for testing regression programs with a test set of data values. In presenting the test data, he notes, for example, that if an explanatory variate, e.g., $X_5 = 2.1$ is used for all observations, when it is also used as a constant variate, depending upon the program, it can lead to "a run-time erorr, or to nonsensical results" or even program failure.(4)

A second area of computational inaccuracy is often found in the algorithm used for computing standard deviation. Several factors contribute to inaccurate results. First, small sets of data with large observation values or observation values differing at the sixth or seventh place can both create problems. Second, the computation of the mean value for the data set may result in a truncated value which then compounds errors as the sum of the squared deviations from the mean is computed. There are two possible errors which result from the two circumstances. The computed standard deviation may be erronenous (which will cumulatively be a source of errors when used in succeeding computation steps such as hypothesis testing) or, more obviously,in some cases, the program will "crash".

One simple test of accuracy is to compute the mean and standard deviation for the following set of values:

888.889, 888.888, 888.887

The rounding error in some programs will cause them to "crash" with this task as a negative value may be obtained.

A second test is to hand-calculate the standard deviation for a raw data set using at least five significant figures if available, then compare the hand results to the computer's.

The third area of numerical accuracy problems extends beyond any one statistical procedure to apply to all results of the computations within a package. Accuracy of results depends upon the input data and, also, the number of output digits. Typically, an error occurs when large variable values are truncated within computation steps. In testing the six microcomputer statistical packages, Lachenbruch's standard of comparison was to 15 significant figures. He also compared two main-frame packages, SAS and GLIM, which each computed to at least 5 significant figures. A suggested standard might be for accuracy to a minimum of five significant figures for statistical results.

The problems and tests which have been outlined provide some suggestions for selection criteria for accuracy in a statistical package. Often, of course, one cannot first test a program, more frequently, evalution is based on product literature and reviews. Many software publishers and reviews will state the precision of calculation results for the package. One microcomputer author, Wilkinson, has been especially concerned with main-frame computing accuracy, and has created a microcomputer package, SYSTAT, which performs calculations in double precision with 15-digit accuracy.(5) In some cases, the Longley data test results are also available in reviews or the product literature. The authors of CRISP, another microcomputer package, cite good test results for the product using the Longley data.

2. STATISTICAL TESTS AND LEVELS OF MEASUREMENT

Microcomputer statistical packages differ in the statistical procedures which are included. In survey research the level of measurement is frequently at the nominal level. Some ordinal measures with ranking as well as interval data might be included. In order to appropriately handle the different levels of measurement in the analysis of survey data, the researcher should consider the statistical procedures offered by various packages.

For nominal data,frequencies, histograms, Chi-sqare,and calculation of lambda are useful. For ordinal data, procedures for calculating Kendall's tau, the most commonly used ordinal measure of assocation, as well as Spearman's rho for rank-order correlations should be included.

Interval data analysis will require descriptive statistics, t-tests, correlation, ANOVA, regression (including multiple stepwise regression) crosstabulations,(including chi-square, and other bivariate measures), multivariate factor analysis, and possibily, time-series analysis. Nonparametric tests in addition to the ordinal measures listed above may also be required. There are considerable differences between statistical packages in the availability of nonparametric tests. In a review of 24 microcomputer statistical packages (Carpenter, et al 1984), 13 did not include any nonparametric tests and only four of the twenty-four reviewed, MICROSTAT, NUMBER CRUNCHER, NWA STATPAK, and STATPRO, included more than two nonparametric tests. (6)

3. FILE HANDLING

The utility programs for file building and file handling are other technical criteria especially important in statistical packages. File handling criteria to consider include: file size limitations, ease of file building, data handling tasks such as, searching, sorting, and transformations, and file output.

Some packages are limited in the size of the files, for example, there may be limitations in the number of observations or variables which can be accomodated. One package we reviewed, TELOFACTS, which is designed specifically for survey research, was limited to variables with no more than five values.

Ease of data entry for file building is impossible to evaluate from product literature. A review which describes data entry procedures is likewise a poor substitute for actually testing the data entry procedures. In some packages there are cumbersome and time-consuming procedures required for correcting data entry errors. A program which has a feature in the file building which prompts for variables by name, rather than variable number, for example, facilitates data entry and reduces mechanical errors.

A major consideration for researchers with existing data files is the ability of the package to read ASCII or DIF format files into its file building program. Many packages do not have the capacity to import or export data files. As an example of a statistical package which can import files, CRISP, allows direct transfer of DIF-format files from 1-2-3, Visicalc, dBII and dBIII, and Rbase 5000. CRISP also reads ASCII files which can be downloaded from a mainframe or read from an ASCII file created by another program.

Other data handling features to assess are the procedures available for searching and sorting by variables or cases, and transformations. In a review of four packages, CRISP, SPSS PC, STATPRO, and SYSTAT, (Fridlund 1985) the features were listed as available in each. A qualitatitive review of ABSTAT, CRISP, MICROSTAT, SL-MICRO, AND SYSTAT, (Goodban and Hakuta 1984) rated the five on those data handling features and found variation in the packages. For example, in rating the utilites to sort and merge files and select cases, CRISP and MICROSTAT were given an "excellent" rating with "good" ratings for ABSTAT and SYSTAT and no rating for SL-MICRO.(7)

File output criteria to consider are readability and support of output devices. The variable and value labelling features available control readability. A package which includes short and extended labels for variables as well as value labels provides a good level of readability for the output. Another output feature to consider is the ability to write output to a disk file as well as the support of printer and plotter devices.

B.GENERAL CRITERIA

1. USER COMPUTING ENVIRONMENT

If more than one type of microcomputer system is available, which is often the case in an academic setting with offices and labs equipped with several types, the user computing environment will not be as important as a general criteria for selecting the statistical package. However, the search for software usually is delimited by one or two systems and narrows the selection of packages to those which match the computing environment (hardware) of the user.

Software packages have a variety of hardware requirements: some are system specific, e.g., IBM-PC, APPLE, or MacINTOSH, DEC RAINBOW, etc., which are related to the operating systems available in a system (PC-DOS, Apple DOS, CP/M, etc.) Another requirement is minimum RAM (Random Acess Memory) which ranges from 48K to 500K for the statistical packages currently available. For most of the larger statistical packages, 128K or 256K is required. The number of disk drives or a hard disk requirement also varies from package to package. To increase the speed of processing with a statistical package, an additional math chip (such as the 8086 or 8087) may be required. Finally, for the packages which support graphics and plotters, specialized hardware is necessary.

Another consideration in software selection is cost, and the cost of software is related to the computing environment. In some instances, program requirements for memory or a math chip will be costs in addition to the software package price. There is a wide range in costs, from under $100 to $1,995, for microcomputer statistical packages. In rating the packages, a cost code for three categories is useful. The lowest range is for packages under $200, a middle range for packages between $200 - $500, and an upper range of those priced over $500.

2. EASE OF USE

While ease of use will obviously vary by user depending upon the individual's previous experience with microcomputing and statistics, there are several key factors in evaluating ease of use. First, are the programs menu-driven or command-driven? Menu-driven programs prompt the user to select options for the execution of the program rather than type commands for the next steps in program execution. as is the case in command-driven programs. In order to work with a command-driven program, the user must be quite familiar with the syntax (rules) of the commands. A missing blank space or comma, or other minor entry difference can create an error message and fail to move the program to the next step. In contrast, in a well-constructed menu-driven program, the user is only required to enter the number or letter(s) to designate the choice of a program step.

Main-frame statistical packages such as SPSS and SAS are command-driven as is the microcomputer version of SPSS. SYSTAT and STATPAC are also command-driven. STATPAC emphasizes in its literature that it is primarily intended for professionals. (A phone inquiry we made provided an estimate of one month to learn the programs.) SYSTAT is described as "cordial" rather than "user friendly" by its publisher. (8)

Good documentation, especially a tutorial and a manual which is instructive rather than technical-descriptive, is critical for ease of use. In some cases, the manual documentation is not adequate for first-time users. For example, the manual for MICROSTAT gave insufficient instructions to get the program running in two different dealer attempts and a complete demonstration proved impossible at that time.

The documentation of a statistical package should include a technical section which provides a description of the computational formulae used.

A third evaluation criteria for ease of use is the availability of good user support, such as an 800- telephone "hot line."

As a practical matter in rating ease of use, determine the amount of disk swapping required (if a hard disk is not used) for operating the package. The larger packages require numerous program disks. For example,STATPRO includes 20 disks and SPSS has 9 disks. With multi-program disks, transition between data analysis tasks is slower as different statistical procedures are often on separate disks.

We conducted our search for a statistical package using the research requirements, technical and general criteria presented in the above section. (To aid the reader in selecting a package, a four page reference guide to statistical software begins on page 22.) After an extensive search which included testing a number of packages or, in some cases, the demonstration disks, we narrowed our choice to two packages, STATPRO and CRISP. STATPRO met all of our requirements except for cost criterion. Even with a 50% educational discount which lowered the price to $997.50 it was above the medium range level, and we selected CRISP which was priced below $500.

CRISP met our criteria in all areas. Technical criteria of accuracy, availability of statistical procedures and file handling were readily satisfied. The authors of CRISP document their testing of the Longley data and statistical routines are accurate to six significant figures. A comprehensive package of statistical routines including nonparametric tests are available. Within the statistical programs: frequency, crosstabulation ,descriptive statistics, t-tests, ANOVA, crosstabulation, multiple stepwise regression, correlation, principal components factor analysis, nonparametric rank tests, and parametric equivalents, there are extensive tests and routines included. The ANOVA program, in particular, is exceptional in its offerings.

File handling features were a particular deciding factor in our selection of CRISP. File size is virtually unlimited. There is no limitation on the N of observations and up to 252 variables may be used. File handling utility programs include an easy file building program which accepts data from other ASCII or DIF files. The file management tasks such as searching, sorting and transformation tasks are performed easily by the utility programs.

General criteria met by CRISP included our computing environment specifications for a program for an IBM-PC, 256K, with 2 disk drives. Ease of use criteria are readily met. (9) It is menu-driven and displays many useful prompts and error messages. There is good documentation in a 300 page manual. In the manual a sample data set is provided and used for all examples, thus providing continuity in documenting the use and interpretation of various statistical routines in CRISP. Finally, the telephone inquiries are given prompt and authoritative responses.

III. USING A MICROCOMPUTER PACKAGE

The use of the CRISP package for handling our ten-year study data provides an example of the comparative advantages of using a microcomputer rather than a main-frame statistical package. Our experience with data entry and data analysis using CRISP is described first, then a comparison of time and cost factors for the two methods are discussed.

A. STUDY DATA DESCRIPTION

Our survey instrument of 53 variables replicates (verbatim) the original Baumhart instrument (Baumhart 1961) and includes an additional 12 variables related to student attitudes towards their ethical decisions, especially attitudes and reported behavior regarding student cheating. The measures consist of nominal, ordinal and interval data. The methodology, as described above, was not based on random sampling, rather, student populations in business courses were surveyed. Therefore, our statistical analysis design relies less on inferential statistical procedures and more on nonparametric measures.

In addition to our own data, the comparative data of the two previous studies required tests for significant differences. The raw data for the early studies was unavailable to create files of observations. In order to complete a comparative analysis, therefore, an interactive statistical procedure was necessary.

There are two major files created by the study data: 1) the Master data file (N=1,274) for the 1975-1985 observations collected from 36 sections of undergraduate and 13 sections of graduate students enrolled in selected Business courses at universities in five geographic regions; 2) the National Study data file for the 1,500 observations collected at 20 institutions in five geographic regions during Fall semester 1985.

B. FILE HANDLING TASKS USING THE MICRCOMPUTER PACKAGE

Data entry for the Master file was by keyboard. The file building procedures using CRISP were as follows:
A file containing 0 observations but with complete labels for variables and values was built and then used as a type of template for all study files. This eliminated the necessity of repititious variable labelling entries each time a file was created. The label file was copied for each file,and then observations were entered. Each data set, an individual class, was entered as a separate file. Data analysis was performed by individual class, followed by concatenation of the files by year, and a final concatenation of the ten years' data for the Master file.

The utility programs for file building and concatenating were easy to use with menus and good documentation. The ease of use of the programs coupled with the portability of the data disks permitted data entry at multiple sites, on individualized schedules, and using non-technical student support staff on occasion.

Another mode of data entry using the main-frame system for downloading of files was tested with four files prior to the national study. The optical scan reader was used to read and create files in the main-frame system, then using a modem the files were downloaded. A useful utility program for the formatting of the ASCII files was available in CRISP. Numerous errors in the files, however, due to main-frame system errors in the optical scan process, led us to abandon the procedure. It was determined that data entry by keyboard was less time-consuming than the downloading procedure. When the national data was entered, student assistants averaged approximately 60 instruments (with 53 variables) per hour.

In another project using CRISP, a third method of data entry was used successfully. Data on disk, in ASCII format, was read into the file building program after creating a format file. Once created, a format file can be saved in CRISP and reused for building additional files from disk or downloading.

C. DATA ANALYSIS USING CRISP

Our initial data analysis runs used the statistical program of Frequency Distributions which calculates absolute and relative frequencies and cumulative frequencies, measures of central tendency and dispersion. Missing data is handled separately (see Table 1). The program has the option of displaying an histogram for the distribution (see Figure 1). The histogram permits choice of treating the variable as nominal, ordinal or continuous.

The responses for the variable OBTEXAM shown in the output examples refer to question #53 of the instrument: "If you knew where and how to get an exam before it was to be given and there was little risk of being caught, would you: 1) try to get it? 2) get it and not share it? 3) get it and share it with your close friends? 4) get it and try to sell it? or 5) advise the teacher of the test's vulnerability?

Table I

Frequency distribution for OBTEXAM

Value	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0	353	353	27.71	27.71
1	133	486	10.44	38.15
2	51	537	4.00	42.15
3	204	741	16.01	58.16
4	9	750	0.71	58.87
5	524	1274	41.13	100.00

Number missing (9) = 0 (0.00 percent of observations)

Statistics on 1274 observations with non-missing data:

Mean =	2.750	Standard deviation=	2.132
Median=	3.000	Variance=	4.544

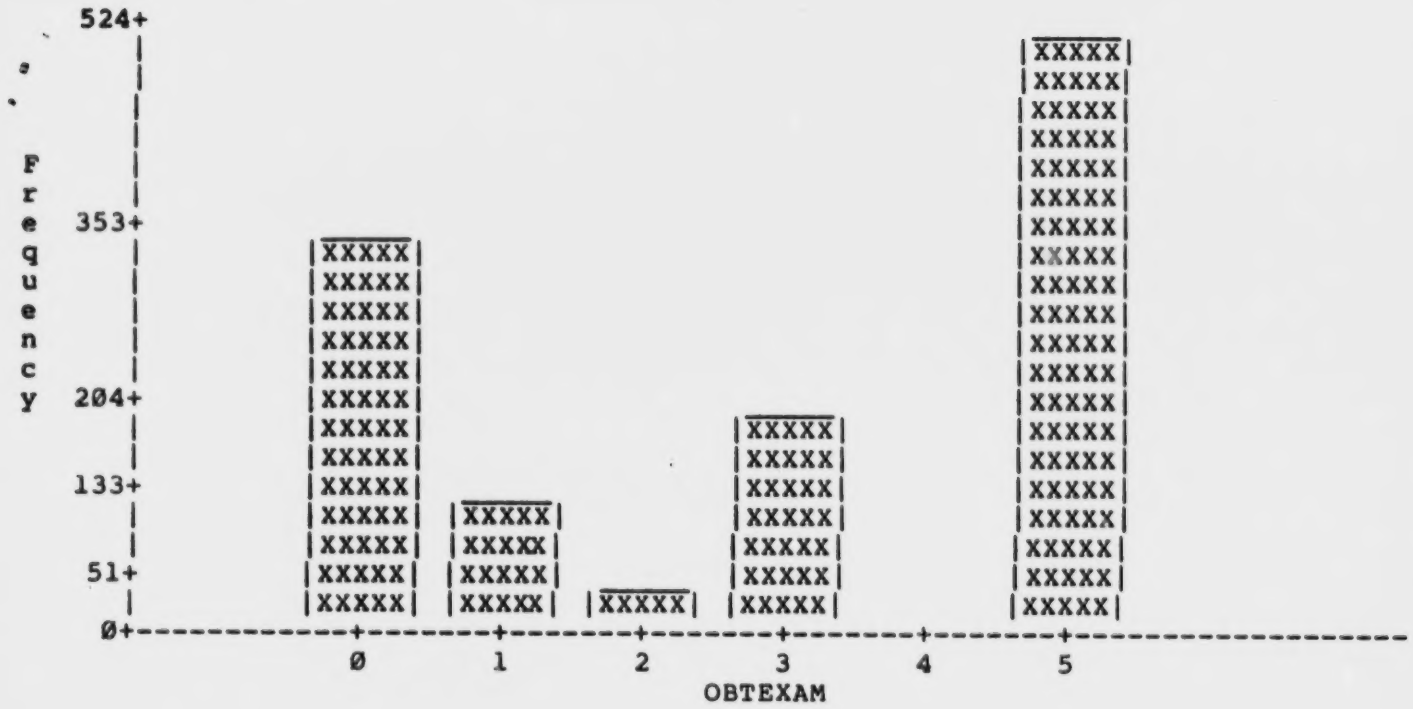

Figure 1

The TABLE program which performs crosstabulations with recodes was used extensively in the data analysis. There are several features that made the program especially useful for our data analysis needs. First, TABLE is the only CRISP program which allows interactive data entry for a statistical routine. The other programs in the package require defining a data file which has been built previously. The special requirement in our research was to be able to process summary values for the two previous studies as the raw data was unavailable. TABLE permitted entering row and column values within the program for the comparative data and then used the Master file data for the values for our study. We were thus able to run tests of association for the three groups.

Another special feature of TABLE is the ability to recode columns or rows to collapse data within the same run. It is possible to run a table, view the output on the screen and immediately collapse rows and/or columns and instantly view the collapsed table. This feature is especially useful in cases where the expected frequencies in a cell is <5 when computing Chi-square, and the data categories permit collapsing. Examples of two tables for the variable OBTEXAM crosstabulated by SEX are presented below in Tables 2 and 3.

Table 2

Tabulation of SEX (rows) by OBTEXAM (columns) Original Table
(Frequency/Row percent/Column percent)

	0	1	2	3	4	5	
0	12	3	4	7	0	16	42
	28.57	7.14	9.52	16.67	0.00	38.10	3.30
	3.40	2.26	7.84	3.43	0.00	3.05	
1	201	82	29	141	7	258	718
	27.99	11.42	4.04	19.64	0.97	35.93	56.36
	56.94	61.65	56.86	69.12	77.78	49.24	
2	140	48	18	56	2	250	514
	27.24	9.34	3.50	10.89	0.39	48.64	40.35
	39.66	36.09	35.29	27.45	22.22	47.71	
	353	133	51	204	9	524	1274
	27.71	10.44	4.00	16.01	0.71	41.13	

Statistics for table of SEX by OBTEXAM

Chi-square (10 df)	= 33.2191	(P<0.0003)
WARNING: P-value may not be accurate		
Minimum expected frequency is 0.297		
4 cells have expected frequencies less than 5.		
Cramer's V	= 0.1142	
Contingency coefficient	= 0.1594	
Kendall's tau b	= 0.0678	(P<0.0074)
Kendall's tau c	= 0.0620	(P<0.0074)
Kruskal-Wallis H with SEX dependent	= 26.4845	(P<0.0001)
Kruskal-Wallis H with OBTEXAM dependent	= 980.3889	(P<0.0000)
Lambda with SEX dependent	= 0.0000	
Lambda with OBTEXAM dependent	= 0.0000	
Symmetric lambda	= 0.0000	
Uncertainty coeff. with SEX dependent	= 0.0163	
Uncertainty coeff. with OBTEXAM dependent	= 0.0092	
Symmetric uncertainty coefficient	= 0.0118	
Pearson's r (1272 df)	= 0.0675	(P<0.0160)
Kappa (% Agreement= 8.79 Chance= 8.41)	= 0.0041	(P<0.5085)

Odds ratio not defined when table is not 2 x 2.
Fisher's exact test not defined when table is not 2 x 2.

Table 3

Tabulation of SEX (rows) by OBTEXAM (columns) Collapsed Table
(Frequency/Row percent/Column percent)

	A	B	C	
0	15	11	16	42
	35.71	26.19	38.10	3.30
	3.09	4.31	3.00	
1	283	170	265	718
	39.42	23.68	36.91	56.36
	58.23	66.67	49.72	
2	188	74	252	514
	36.58	14.40	49.03	40.35
	38.68	29.02	47.28	
	486	255	533	1274
	38.15	20.02	41.84	

Statistics for table of SEX by OBTEXAM

Chi-square (4 df)	=	25.0778	(P<0.0000)
Cramer's V	=	0.0992	
Contingency coefficient	=	0.1389	
Kendall's tau b	=	0.0732	(P<0.0052)
Kendall's tau c	=	0.0632	(P<0.0052)
Kruskal-Wallis H with SEX dependent	=	23.9633	(P<0.0000)
Kruskal-Wallis H with OBTEXAM dependent	=	8.7162	(P<0.0128)
Lambda with SEX dependent	=	0.0000	
Lambda with OBTEXAM dependent	=	0.0243	
Symmetric lambda	=	0.0139	
Uncertainty coeff. with SEX dependent	=	0.0125	
Uncertainty coeff. with OBTEXAM dependent	=	0.0095	
Symmetric uncertainty coefficient	=	0.0108	

Pearson's r not available for collapsed tables.
Kappa not available for collapsed tables.
Odds ratio not defined when table is not 2 x 2.
Fisher's exact test not defined when table is not 2 x 2.

After viewing the collapsed table, the original table can easily be restored simply by selecting that option in the menu (without waiting for recomputation.) There is a good selection of statistics which can be computed for the table. The examples above displayed all of the options. A submenu provides choice of 11 contingency table statistics thus permitting display of only relevant statistics for the data set.

Throughout all of the statistical programs in CRISP, the prob-value is computed and displayed. The TABLE examples above show the listings with prob-values.

The program CROSSTAB in the CRISP package will also compute the same statistics but will not permit recoding. CROSSTAB prompts for selection between two methods for generating tables before entering variables. The screen display is quoted below:

> "Two way crosstabulations are generated for pairs of variables in two lists. Two methods control the number of tables generated. Method 1 generates a table for each possible pair of list 1 variable with list 2 variable. Method 2 generates a table for each corresponding variable pair in the two lists. Enter 1 for method 1 or 2 for method 2:"

It is worth noting, for large data files, that CROSSTAB will allow processing of as many as 300 tables in one pass through the data.

D. COMPARATIVE ADVANTAGES

The output and descriptions of programs given above provide examples for statistical routines common in all survey research. The selection of statistical procedures in CRISP is extensive and can meet the needs of any survey research design. The availability of comprehensive statistical packages combined with the flexibility of microcomputing suggests a comparative advantage for using a microcomputer rather than a main-frame package.

A comparison between the use of a microcomputer or a main-frame package can be summarized for three main areas: differences in costs, access and time factors.

COMPARATIVE COSTS

MICROCOMPUTER	MAIN-FRAME
DATA STORAGE	
Data Disk <$2 no storage costs	On-Line: CPU costs for system range: $.62/second - $1.00/hr. Tape:storage,mounting fees avg.$3/month, $2/mount Disk Storage: (by track and block) avg. $.0085/KB/month Cards: no storage costs
DATA ENTRY	
hourly wage for routine clerical keyboard entry: $<5/hr.	trained personnel for time-share system: requires knowledge of operating system, file building, statistical package: $5 -$10/hr. keypunch:avg. $10/hr.
DATA PROCESSING	
INTERACTIVE MODE: research staff time costs	INTERACTIVE MODE: computer time costs at highest CPU rates; research staff time costs BATCH PROCESSING: computer time costs at lower rates available; research staff time costs
DATA OUTPUT	
minimal paper costs: $25/1000 = $.025/page	Print costs depend on type of printer, time of printing, and type of paper. Cost range averages $.04 - .08/page. Special forms cost more and the use of plotters is costly ($1/plot and $1/minute)
OVERHEAD and CAPITAL COSTS	
original capital costs:$2,000 - $6,000	Time-share fees,built-in overhead, plus high capital costs.

Cost estimates based on system costs at a major research university

COMPARATIVE ACCESS AND TIME FACTORS

MICROCOMPUTER	MAIN-FRAME
ACCESS	
flexibile, transport-ability of program and data disks, permits access at convenient times and locations	inflexible, terminal locations in labs and offices can restrict access
TIME FACTORS	
INTERACTIVE MODE: immediate output to screen or printer	INTERACTIVE MODE: (high cost time period) slow turnaround for printed output in time-share system
no break between initial analyses and further in-depth analysis	typical to separate initial data runs from in-depth runs
no downtime	unscheduled downtime can occur at any time; routine schedule for maintenance requires weekly downtime
	BATCH PROCESSING: slow turnaround - hours of delay in obtaining print-outs

A comparison of features suggests a strong advantage in utilizing a microcomputer rather than a main-frame system for efficiency and low cost in survey research. There are times, of course, when the differences between the two types of systems might be considered a false dichotomy and a symbiotic relationship can prove effective for a particular research situation. At one major medical research center using CRISP, for example, extensive data files are maintained in auxiliary storage on the main-frame. Analyses of subsets of data are downloaded for the microcomputer analysis, while large data sets are sometimes processed on the mainframe due to memory constraints of the microcomputer.

IV. THE IMPACT OF CHANGING TECHNOLOGY ON SURVEY RESEARCH

Within the past several years technological innovations in the development of microcomputers, especially the expansion of memory and speed of processing has changed the alternatives for data processing. Just as the introduction in the 1960's of main-frame computing changed the scope and feasibility of conducting social science research, the introduction in the mid-1980's of powerful microcomputer tools will change the methods in conducting survey research.

Efficiencies in time and cost factors will enable more researchers to engage in data collection and analysis. As costs decrease for data processing, and particularly as the cost basis for data processing become less related to costs/observation, the size of samples can also increase. Computing statistical tools which are only available on a more restricted basis in the main-frame systems are now accessible to any researcher who has access to a low-cost microcomputer and software.

The change in technology can influence future survey research methodology. Several changes can be predicted. The ease of data storage, for example, suggests incorporating more pre-post tests in educational research. Another change will be in sampling design. Random sample selection and optimum sample size can be calculated in seconds using easy microcomputer routines. The ease of sampling selection will encourage broader use of stratified designs. Data analysis models can also become more complex as the benefits of low processing costs and interactive analysis become more widely recognized.

Educational researchers have more powerful tools and freedom with the microcomputer statistical packages. The challenge for survey research is to keep up with technology!

STATISTICAL SOFTWARE REFERENCES

STATISTICAL PACKAGE	HARDWARE REQUIREMENTS	COST CODE	COMMENTS
ABSTAT Anderson-Bell POB 191 Canon City,CO 81212 (303) 275-1661	IBM-PC or CP/M operating systems 128K 2 Disk Drives	2	Review Codes 4,7 Good Manual
AIDA Action-Research Northwest 11442 Marine View Dr.SW Seattle, WA 98146 (206) 241-1645	Apple DOS 48K 1 Disk Drive	2	Review Codes 4,8
A-Stat Rosen Grandon Assoc. 7807 Whittier St. Tampa, FL 33617	Apple DOS 48-64K depending on system 1 Disk Drive	1	Review Codes 4,8
BMDPC BMDP Statistical Software 1964 Westwood Blvd.#202 Los Angeles, CA 90025 (213) 475-5700	refer to publisher	2/3	PC version of main- frame pkg. Purchase by program
CRISP 2547-22nd Ave. San Francisco CA 94116 (415) 564-7337	IBM-PC/XT/AT/ Compatible 256K 2 Disk Drives	2	Review Codes 6,7
Dynacomp Dynacomp Inc. 1427 Monroe Ave. Rochester, NY 14618 (716) 442-8960	Apple 48K 1 Disk Drive	1	Review Code 4 Purchase by program Multi-pkgs
EPISTAT Tracy L.Gustafson, M.D. 1705 Gattis School Rd. Round Rock, TX 78994	IBM-PC/COMPATIBLE 64K 1 Disk Drive	1*	Review Code 9 PUBLIC DOMAIN
Daisy Rainbow Computing Inc. 19517 Business Center Dr. Northridge, CA 91324	Apple DOS 48K 1 Disk Drive	1	Review Code 8

STATISTICAL SOFTWARE REFERENCES

STATISTICAL PACKAGES	HARDWARE REQUIREMENTS	COST CODE	COMMENTS
GAUSS Applied Technical Systems POB 6487 Kent, WA 98064 (206) 631-6679	IBM-PC 256K 1 Disk Drive 8087 NDP (math chip)	2	fast, accurate
HSD ANOVA, Stats Plus, Regress Human Systems Dynamics 9249 Reseda Blvd., Suite 107 Northridge, CA 91324 (213) 993-8536	Apple DOS 48K 1 Disk Drive	1	Review Codes 4,8 Programs in multi-packages
Introstat Ideal Systems POB 681 Fairfield, IA 52556 (515) 472-4507	Apple DOS, IBM-PC 48-64K	1	Review Code 4
Microstat Ecosoft, Inc. POB Box 68602 Indianapolis, Inc. 46268 (317) 255-6746	IBM-PC or CP/M 128K 2 Disk Drives	2	Review Codes 4,7
Micro-TSP College Division McGraw-Hill Princeton Road Hightstown, NJ 08520 (609) 426-5000	Apple or IBM-PC DOS 48-64K 1 Disk Drive	2	Review Code 4
Number Cruncher Dr. Jerry Hintze 865 East 400 North Kaysville, UT 84037 (801) 546-0445	IBM-PC/XT, CP/M 64K 1 Disk Drive	1	Review Code 4
NWA Statpak Northwest Analytical 520 NW Davis St. Portland, Oregon 97305 (503) 224-7727	IBM-PC, CP/M 48K 2 Disk Drives	2	Review Code 4

STATISTICAL SOFTWARE REFERENCES

STATISTICAL PACKAGE	HARDWARE REQUIREMENTS	COST CODE	COMMENTS
P-STAT P-STAT,Inc. 471 Wall St. POB AH Princeton, NJ 08542 (609) 924-9100	IBM-PC/XT/AT/ Compatibles 256K 2 Disk Drives	2	PC version of a main-frame pkg.
SpeedSTAT 1,2,and 3 SoftCORP Intern'tl.,Inc. 229 Huber Village Blvd. Westerville, OH 43081 (513) 891-5044	Apple DOS 48K 2 Disk Drives	2	Review Code 4 Programs in 3 pkgs.
SL-Micro Questionnaire Service Co. POB 23056 Lansing, MI 48909 (517) 641-4428	IBM-PC DOS 128K 2 Disk Drives	2	Review Code 7
SPSS PC SPSS Inc. 444 N. Michigan Ave. Suite 3000 Chicago, IL 60611 (312) 329-2400	IBM-PC XT, COMPATIBLES running PC or MS-DOS 2.0 (or later) 320K, at least 3 megabytes of hard disk space/1 floppy Disk Drive	3	Review Code 6 PC version of major main-frame package
Statpac Walonick Assoc. 5624 Girard Ave.,So. Minneapolis, MN 55419 (619) 866-9022	IBM-PC,Kaypro, Radio Shack memory requirement depends upon systems configuration	2	Review Code 4 similar to SPSS main-frame pkg.
Statpro Wadsworth Electronic Publishing Co. 20 Park Plaza Boston, MA 02116 (617) 423-0420	Apple DOS or IBM-PC memory requirement depends upon systems config. 2 Disk Drives	3	Review Code 4,6
The Statistician Quant Systems Box 628 Charlestown,SC 29402 (803) 571-2825	IBM-PC, CP/M Radio Shack TRS-80 Kaypro, Xenix memory requirement depends upon system	2	several versions available

STATISTICAL SOFTWARE REFERENCES

STATISTICAL PACKAGE	HARDWARE REQUIREMENTS	COST CODE	COMMENTS
Systat Systat,Inc. 1127 Asbury Ave. Evanston, IL 60202 (312) 864-5670	IBM-PC DOS 256K 2 Disk Drives	2	similar to SAS, a main-frame pkg review codes 4,6, 7 and 9
TELOFACTS I, II dilithium Press 8285 SW Nimbus Beaverton,Oregon 97005 (800) 547-1842	Apple DOS, TRS-80 memory requirements depend upon systems configuration	1	package for survey research: design and analysis-(nominal data)
TWG/ELF, TWG/ARIMA The Winchendon Group 3907 Lakota Rd. POB 10339 Alexandria,VA 22310 (703) 960-2587	Apple DOS, IBM-PC, CP/M,(depending upon package) memory requirements depend upon systems configuration	2	two pkgs: spec.stats review code 4

COST CODE: 1 = UNDER $200
2 = $200-500
3 = OVER $500

REVIEW CODE: See REFERENCES (e.g., CARPENTER = 4)

N.B. The data listed above is subject to change. Verify specific system requirements and costs by contacting the publishers directly.

REFERENCES

1. Bannister, Hank. "Micro-Applications Broaden the Impact of Statistical Analysis." InfoWorld, October 14,1985. pp. 28-30.

2. Baumhart, Raymond C. "How Ethical are Businessmen?" Harvard Business Review, July-August 1961. pp.7ff.

3. Brenner, Steven and Molander, Earl A. "Is the Ethics of Business Changing?." Harvard Business Review, January-February 1977, pp. 57-71.

4. Carpenter, James; Deloria, Dennis; and Morganstein, David. "Statistical Software for Microcomputers." BYTE, April 1984, pp.234-264.

5. Chambers, J. Computational Methods for Data Analysis. New York: John Wiley and Sons, 1977.

6. Fridlund, Alan. "Taking the Numbers Bull By the Horns." InfoWorld, February 11, 1985, pp. 42-50.

7. Goodban, Nancy and Hakuta, Kenji, "Statistical Quintet." PC World, August 1984, pp. 186-195.

8. Lachenbruch, Peter A. "Statistical Programs for Microcomputers." BYTE, November 1983, pp. 560-570.

9. Lee, Robert S. and Cohen, Jacob. "Good with Stats: SYSTAT and EPISTAT." PC Magazine, March 5,1985. pp. 179-186.

10. Longley, J.W. "An Appraisal of Least Squares Programs from the Point of View of the User." Journal of the American Statistical Assocation, 1967, volume 62, pp. 819-841.

11. Maindonald, J.H. Statistical Computation. New York: John Wiley and Sons, 1984.

12. Platt, W.G. and Platt, C.A. "Microcomputer Statistical Packages and Regression Performances." American Statistical Assocation 1981 Proceedings of the Statistical Computing Section, pp. 309-316.

NOTES

1. James R. Glenn, Jr. Ph.D., School of Business, Management and Marketing Department, San Francisco State University has been my research colleague in the ten-year survey study.

2. J. Chambers, Computational Methods for Data Analysis, (New York: John Wiley and Sons, 1977).

3. Peter A. Lachenbruch, "Statistical Programs for Microcomputers," BYTE, November 1983, p. 567.

4. J.H. Maindonald, Statistical Computation, (New York: John Wiley and Sons, 1984), pp. 57-58.

5. Robert S. Lee and Jacob Cohen, "Good with Stats: SYSTAT and EPISTAT," PC Magazine, March 5,1985, p. 185.

6. James Carpenter, et al., "Statistical Software for Microcomputers," BYTE, April 1984, p. 246.

7. Nancy Goodban and Kenji Hakuta, "Statistical Quintet" PC World, August 1984, p. 189.

8. Alan J. Fridlund, "Taking the Numbers Bull by the Horns," InfoWorld, February 11, 1985, p. 48.

9. Undergraduate students at San Francisco State University without previous computing experience are using CRISP for courses in statistics and for other coursework requiring statistical analysis.